GEB 6895: Business Intelligence
Department of Economics
College of Business Administration
University of Central Florida
Fall 2019

Final Examination

Wednesday, December 11, 2019 from 6:00 to 8:50 PM
in *your* private mirror of the GEB6895F19 GitHub repo.

Instructions:

Complete this examination within the space on your private mirror of the GEB6895F19 GitHub repo in the folder `final_exam`. Create a folder called `my_answers` that will contain all of your work for this examination. When you are finished, use `git` to `add`, `commit` and `push` your code to your private mirror of the GEB6895F19 repo. To complete this examination, you are free to use any resource available, including books and, most importantly, the Internet. Complete this examination like you would any other project for an employer, except that you cannot discuss the content with your coworkers.

Grading Policy:

- The examination includes THREE questions.
- Question 1 is required and is worth 80% of the grade.
- You may choose either one of Questions 2 or 3 for the remaining 20% of the grade.
- Any progress on the remaining question will count as extra credit toward Question 1.
- Your goal is to demonstrate as many skills as you have, to show an employer what contribution you would make to their team.
- You may not have time to complete all of the sections of Question 1. There are many options to produce the best model you can in the time you have.
- MOST IMPORTANTLY, make sure to `git add`, `commit` and `push` your progress to your private repo before leaving the exam room.

Question 1 Data:

Use the language of your choice to construct a model for used tractor prices with the data from `tractor_sales.csv`. The dataset contains the characteristics and conditions of tractors and includes the following variables.

$saleprice_i$	=	the price paid for tractor i in dollars
age_i	=	the number of years since tractor i was manufactured
$enghours_i$	=	the number of hours of use recorded for tractor i
$johndeere_i$	=	an indicator of whether tractor i is manufactured by John Deere
$spring_i$	=	an indicator of whether tractor i was sold in April or May
$summer_i$	=	an indicator of whether tractor i was sold between June and September
$winter_i$	=	an indicator of whether tractor i was sold between December and March

You also have access to another dataset `tractor_specs.csv`, which an intern has painstakinly compiled from research about the models of tractors that were sold.

$horsepower_i$	=	the horsepower of tractor i
$diesel_i$	=	an indicator of whether tractor i runs on diesel fuel
fwd_i	=	an indicator of whether tractor i has four-wheel drive
$manual_i$	=	an indicator of whether tractor i has a manual transmission

Question 1 Modeling Suggestions (80% of Grade):

Instructions:

Your goal is to estimate a regression model that predicts $saleprice_i$. Regardless of which modeling suggestions that you implement, save the printout of the regression output from your final recommended model, either at the end of your script or in a separate file.

a) First, choose the form of the dependent variable and build an initial model by following these steps:

   i) First, decide whether to build a model to predict the raw variable or to transform it into the log of the sale price instead. Plot a histogram of the dependent variable `saleprice`. If the distribution of sale price is highly skewed, create the new dependent variable.

   ii) Estimate a regression model for your chosen variable that includes all variables in the dataset `tractor_sales.csv`. Which variables explain used tractor prices?

b) Join the datasets `tractor_sales.csv` and `tractor_specs.csv` to obtain the full set of variables (any way you know how). Estimate a regression model that includes all currently available variables. Which variables explain used tractor prices?

c) A used tractor dealer tells you that overpowered used tractors are hard to sell, since they consume more fuel. Tractor prices often increase with horsepower, up to a point, but beyond that they decrease. To incorporate this advice, you create and include a variable for squared horsepower (in addition to the linear horsepower variable).

   i) Hypothesize the signs for the coefficients on horsepower. Perform 1-sided t-tests of the hypotheses for the horsepower coefficients. (That is, check whether the t-statistics are greater than the critical value of 1.645, with the appropriate sign. If so, it is evidence against exclusion of the variable.)

   ii) Compare the model with or without the quadratic functional form for horsepower. Which model do you recommend? Be sure to cite evidence to support your choice.

d) The above datasets did not indicate whether or not the tractors had enclosed cabs, which would be convenient for the operator. Now suppose that a coworker obtained the data by inspecting photographs from the advertisements and saved a list of tractors with cabs in the file `tractors_with_cabs.csv`. Join a new variable to the dataset from `tractor_sales.csv` that indicates whether or not the tractor has an enclosed cab. Estimate the model with the cab indicator included. Should the cab indicator remain in the equation?

e) Now test the joint hypothesis that the time of year has no effect on the sale of tractors. In this test, the null hypothesis is the joint hypothesis that all coefficients on spring, summer and winter are equal to zero. The alternative hypothesis is that one of these coefficients is nonzero.

i) Obtain the Residual Sum of Squares from the unconstrained regression, RSS_U. This is the regression model in which the seasonal indicators are included.

ii) Obtain the Residual Sum of Squares from the constrained regression, RSS_R. This is the regression model in which the seasonal indicators are restricted to zero, i.e. excluded from the regressions.

iii) Calculate the F-statistic for this test. The F-statistic has a value of

$$F = \frac{(RSS_R - RSS_U)/M}{RSS/(N - K - 1)}.$$

where M is the number of restrictions (the number of variables excluded), N is the number of observations and K is the number of explanatory variables in the full model, excluding the constant.

iv) Perform the F-test for the importance of the seasonal indicators at the 5% level of significance. (That is, check that the calculated F-statistic is greater than the critical value of 2.60. If so, it is evidence against the restriction.) Is there evidence that tractor prices follow a seasonal pattern?

Question 2 (Either 20% of Grade or Bonus Points):

Instructions:

Complete all sections of this question for the full credit of 20%.

Your goal is to find the root of the function $f(x) = \log(x) - e^{-x}$ up to a tolerance level of $|f(x)| < \epsilon = 10^{-6}$.

a) Write or modify a function that calculates the root of $f(x)$ using Newton's method. Modify this function to print out, on each iteration, the iteration number and the candidate root of the function.

b) Consider Einstein's method, an improved version of Newton's method, defined by the recurrence relation

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} - \frac{f''(x_i)}{2f'(x_i)} \left[\frac{f(x_i)}{f'(x_i)}\right]^2 . \tag{1}$$

Modify your function for Newton's method to produce a new function that uses the recurrence relation for Einstein's method.

c) Use both functions to compute the root of $f(x) = \log(x) - e^{-x}$ using $x_0 = 2.5$ as a starting point. Observe the candidate roots x_i on each iteration to compare their performance. Which method performs better for this case? Under what conditions would this be the case?

Question 3 (Either 20% of Grade or Bonus Points):

Instructions:

Complete any two of the four parts of this question for the full credit of 20%.

Consider the following nonlinear regression model:

$$y_t = \beta_1 z_t^{\beta_2} + u_t, \tag{2}$$

where u_t is a random error term that is independently and identically distributed with mean zero and (known) variance one. The data are available in the dataset `nl_data.csv` with variables `y` and `z`. Your goal is to estimate the parameters β_1 and β_2 that minimize the sum of squared residuals.

a) Estimate β_1 assuming that you know the true value of β_2 to be 1/2. This can be done either using a standard regression package or by direct calculation. Note that the regression model does not include an intercept coefficient and the error term u_t is added. For reference, the least squares regression estimator without an intercept is

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n} y_t x_t}{\sum_{t=1}^{n} x_t^2}, \tag{3}$$

where x_t is the explanatory variable in the model.

b) Continue under the assumption that β_2 is known to be 1/2.

   i) Create a function `uni_ssr_1(beta_1, beta_2, y, z)` that returns the sum of squared residuals.

   i) Optimize the function `uni_ssr_1(beta_1,...)` to find the optimal value of β_1 under this assumption. Compare it to the estimate from part (a).

c) Now suppose that β_2 is unknown.

   i) Create a new function `uni_ssr_2(beta_2, y, z)` that returns the sum of squared residuals as a function of only β_2. This function should calculate the value of β_1 for a given β_2 supplied as the argument, using the approach in part (a).

   i) Optimize the function `uni_ssr_2(beta_2, ...)` to find the optimal values of β_1 and β_2 using this approach.

d) Continue under the assumption that β_2 is unknown.

   i) Create a function `multi_ssr_1(beta, y, z)` that returns the sum of squared residuals. This function takes in the vector (β_1, β_2) as a single parameter.

   i) Optimize the function `multi_ssr_1(beta, ...)` to find the optimal values of β_1 and β_2 using this approach.